

Exploring how reducing the weight of low exam scores impacts average grades and equity gaps

Nicholas T. Young^{1,*}, Rebecca L. Matz², Eric F. Bell³, and Caitlin Hayward⁴

¹*Department of Physics and Astronomy, University of Georgia, Athens, Georgia, 30605, USA*

²*Center for Academic Innovation, University of Michigan, Ann Arbor, Michigan, 48104, USA*

³*Department of Astronomy, University of Michigan, Ann Arbor, Michigan, 48105, USA*

⁴*Institute for Social Research, University of Michigan, Ann Arbor, Michigan, 48104, USA*



(Received 13 August 2025; accepted 26 June 2026; published 4 August 2026)

Students interpret grades as signals of their strengths, and grades inform students' decisions about majors and courses of study. Differences in grades that are not related to learning can impact this judgment and have real-world impact on course-taking and careers. Existing work has examined how an overemphasis on high-stakes exams can create equity gaps where female students and Black, Hispanic, and Native American students earn lower grades than male students and Asian and white students, respectively. Yet, minimal work has examined how the weighting of individual midterm exam scores can also contribute to equity gaps. In this study, we examine how three midterm exam score aggregation methods for final grades affect equity gaps. We collected midterm exam data from approximately 6000 students in an introductory physics course over 6 years at a large, research-intensive university. Using this dataset, we applied common midterm exam score aggregation methods to determine their impact on aggregated midterm exam grades: dropping the lowest midterm exam score, replacing the lowest midterm exam score with the final exam score if higher, and counting the highest midterm exam score more in the final grade calculation than the lowest midterm exam score. We find that dropping the lowest midterm exam score resulted in the largest increase in final grades, with students with lower grades benefiting the most. However, we find no evidence that these alternative midterm exam aggregation methods close equity gaps. Instead, we find some evidence that such practices could increase equity gaps between male and female students, though the effect is very small in relation to the overall grade increase. Implementing the alternative midterm exam aggregation methods examined here may be useful for instructors wanting to raise course grades or give lower-scoring students a boost. However, they do not appear to be effective in reducing equity gaps.

DOI: [10.1103/ym5f-hs2g](https://doi.org/10.1103/ym5f-hs2g)

I. INTRODUCTION

Introductory courses have a large impact on undergraduate student retention [1]. Final course grades in these courses may be interpreted by students to reflect what they are good at and can inform their choice to remain in or change their major [2–4].

Yet, grades are not assigned equally across different groups of students, even when controlling for prior performance or academic ability [5]. For example, a multi-institutional study found that men tended to earn higher final course grades than women in introductory physics relative to their other courses [6]. Other studies found that men tended to earn higher final course grades than

women across the physics curriculum [7], and middle- and upper-class, continuing-generation, white, and Asian students tended to earn higher final course grades than students of color, first-generation students, and low-income students [8]. These grade differences affect who earns a science, technology, engineering, and mathematics (STEM) degree, as prior work has found that earning low final course grades early on in STEM courses decreases the chance that a student will graduate with a major in STEM [1,9,10].

In addition, the effect is not the same for all students. Studies have found that women who failed a key “weed-out” STEM course were less likely to graduate with a STEM degree than men who failed the same course [11]. More broadly, men appear to be less sensitive to feedback about their actual course performance and hence remain in STEM fields despite having lower grades than women who leave STEM [12]. Further, men have significantly higher levels of self-efficacy than women, even when their actual performance is similar [13]. Other work has found that the relationship between earning a lower final course grade in

*Contact author: nicholas.young@uga.edu

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

introductory STEM courses and not earning a STEM degree was stronger for students from minoritized races and ethnicities compared to students from majoritized races and ethnicities [14].

As physics remains one of the least diverse STEM fields in terms of sex, race, and ethnicity [15], various studies have focused on how grading approaches in physics, particularly how final course grades are determined, could be more equitable and contribute to a more equitable education environment. For example, Webb *et al.* found that grading students directly on the 4.0 scale rather than the percent scale, both on individual exam items and for the final course grade, significantly reduced the number of final course grades less than a C— [16] and later work found that the same grading system could reduce equity gaps by 20%–25% [17]. Other work has found that reducing the weight of high-stakes exams in the final course grade would eliminate the male-female grade gap and also reduce equity gaps for underrepresented minority students [18], suggesting that grade component weight, in addition to the grading system, is important for demographic equity. In a similar line of work, Webb and Paul found that allowing retakes of high-stakes exams removed equity gaps for women in introductory physics [19] and that the gaps disappeared because the stakes of each exam were reduced, not because women were more likely than men to retake exams or because they used the retake as an opportunity to better learn the material [20].

Here, we seek to add to these studies by conducting a study of how altering the stakes of individual midterm exams, rather than the overall weighting of midterm exams relative to other graded components, may affect students' final course grades, including equity gaps. We specifically examine the impact of three midterm exam aggregation methods compared to the equal weighting of each midterm exam common in physics courses (e.g., [18]): dropping the lowest midterm exam grade, replacing the lowest midterm exam grade with the final exam grade if it is higher, and weighting the best midterm exam score more than the worst midterm exam score.

These techniques rest on the assumption that students show variations in midterm exam performance over time, such that these alternative aggregation methods could make a meaningful difference in aggregated student final course grades. Previous work in a single course at a single institution suggests that this is likely the case [21]. Our first research question (RQ) then seeks to confirm that this assumption is met:

1. How do student grades vary across midterm exams within an introductory physics course?

Assuming that student midterm exam grades vary over the semester, we then ask:

2. What impact do alternative aggregation methods have on students' overall midterm exam grades as compared to weighting each midterm exam equally?

3. How do the impacts of the alternative aggregation methods vary in terms of the students' final grade?

Finally, we then expand the second research question to consider the potential impacts of these alternative aggregation methods on equity gaps:

4. How do the impacts of the alternative aggregation methods vary based on the students' sex or race?

For research questions 2, 3, and 4, we make the critical assumption that student behavior would not change if these alternative aggregation schemes were implemented, an assumption made in similar studies (e.g., [18]), and one we discuss further in Sec. V.

The rest of the paper is organized as follows: In Sec. II, we describe the course deficit model that underlies this work as well as how we are conceptualizing equity to answer research question 4. In Sec. III, we describe how we obtained the data for this study, the alternative aggregation methods, how we created the demographic groups, and our mixed-effects model. In Sec. IV, we show that students exhibit variation in midterm exam grades throughout the semester and that while alternative aggregation methods increase grades on average, they do not close equity gaps. In Sec. V, we provide explicit answers to our research questions and describe the implications of these findings. In Sec. VI, we discuss the limitations of this study and how future work may address these, including conducting studies in courses that implement these alternative aggregation methods.

II. RESEARCH FRAMING

We use two frameworks to guide our answer to research question 4. First, we use the course deficit model to argue that observed equity gaps in a course are the result of the structures of the course rather than an underlying deficiency of one group compared to another [19,22]. That is, course characteristics such as grading policies, content delivery, resources provided, and instructor actions are the cause of equity gaps between different groups of students. Under this paradigm, the solution to closing equity gaps is to alter course structures, rather than to seek to alter student behavior.

Second, because there are multiple definitions of equity [23,24], we consider two definitions of equity used within physics education research: equity of parity and equity of fairness [25]. Equity of parity means that all groups experience the same outcome on average. In the context of this study, if all student groups score, on average, the same on midterm exams after implementing one of the alternative aggregation methods, we would say that method achieved equity of parity because the outcome is the same for all groups, even if there were initial differences. For equity of parity to occur, groups that initially score lower must benefit more from the intervention in order to achieve an equal outcome. This is the common

definition of equity used in previous studies when referring to equity gaps (e.g., [8,26,27]).

In contrast, equity of fairness requires that all groups receive, on average, the same benefit. If all student groups increase their score by the same amount on average after implementing one of the alternative midterm exam aggregation methods, we would say that method achieved equity of fairness. Because all groups receive the same benefit, any existing performance differences between groups before the intervention are maintained after the intervention.

III. METHODOLOGY

A. Data collection

With exemption from review by our Institutional Review Board, data for this study come from ECoach users enrolled in a first-semester, calculus-based introductory physics course (referred to as Physics I) between Fall 2017 and Fall 2023, excluding the Winter 2020 semester when courses and course grading policies were most significantly impacted by the COVID-19 pandemic. ECoach is a free, personalized, web-based coaching tool offered to students in introductory STEM courses (including Physics I) at the University of Michigan [28,29]. Student use is optional in the course under study, but instructors typically incentivize participation by offering a small amount of extra credit, equivalent to an in-class assignment or two, for students who choose to use ECoach.

ECoach provides students with tips on how to approach their course, personalized feedback on their performance, and tools to help them succeed, such as task lists, grade calculators, and midterm exam study playbooks. In order to provide personalized feedback and for the grade calculator to provide accurate estimates, assignment grades for participating students are pulled into the platform from the learning management system, allowing the ECoach gradebook used in this study to act as a duplicate of instructors' official gradebook. Midterm and final exam scores in the gradebook included both the maximum points possible on each exam and the number of points the student earned. The maximum possible score on the midterm exams ranged from 12 to 20 points (median 15 points), and the maximum possible points on the final exams ranged from 25 to 26 points (median 25 points). Students were only able to earn a whole number of points on each exam.

During the time period of study and excluding Winter 2020, there were 60 Physics I sections offered. Of these 60 sections, 30 instructors offered ECoach to their students, while the remaining 30 sections did not offer ECoach. In the 30 sections that did offer ECoach, there were 6082 student enrollments, and 5621 students enrolled in ECoach, resulting in a usage rate of approximately 92%.

TABLE I. The four assessment styles used in the course of interest and the number of students included in the analysis for research questions 2 through 4.

Assessment style	<i>N</i>	Description
3E + F	2290	Physics I sections between Fall 2017 and Fall 2019 had three in-person midterm exams and one final exam
4E + F	1274	Physics I sections between Fall 2022 and Fall 2023 had four in-person midterm exams and a final exam, with the final exam scores recorded in ECoach
4E	750	Physics I sections in Fall 2021 and Winter 2022 had four in-person midterm exams and a final exam, but the instructors did not include final exam grades in ECoach
5E + F	767	Physics I sections in Fall 2020 and Winter 2021 had five online midterm exams and one final exam

Repeat enrollments are included in this count. Approximately 85% of the student enrollments in the course who did not enroll in ECoach withdrew from the course.

Data for this study also came from a simplified version of the university's student records data warehouse [30]. Through the data warehouse, we obtained students' final grades in Physics I, sex, and race and ethnicity.

Due to the variety of assessment styles used by Physics I instructors throughout the study period and recorded in ECoach, we created four separate datasets to correspond to the distinct assessment styles. For the remainder of the paper, we refer to each of these assessment styles by the number of midterm exams offered (E) and whether there was a final exam (F). These styles and a description are shown in Table I.

While the specific weighting of midterm exams and the final varied in each section, the median weighting of the midterm exams was 36% of the final grade (range 35% to 45%), and the median weighting of the final exam was 16% of the final grade (range 15% to 19%). The weight of each midterm exam was equal to the total weight divided by the number of exams. In each section, high-stakes exams accounted for at least half of each student's final grade, with homework, in-class participation, and prelecture activities accounting for the remaining portion of students' final grades.

For each of these four datasets, we apply the same analysis, which is detailed in the next section.

B. Data handling and analysis

1. Grade variations

To answer research question 1, all midterm exam scores were converted to percentages and then to letter grades

TABLE II. Grade ranges for Physics I, ignoring pluses and minuses.

Grade	Percentage range
A	85–100
B	75–84.99
C	60–74.99
D	45–59.99
E	0–44.99

using the grading scale provided by instructors in ECoach. Although the courses used a plus-minus grading system, we used whole letter grades to reduce the number of potential grades from 13 to 5. The percentage-to-grade conversion is shown in Table II. We do this because the exam grades are whole numbers (with a max score of 12, 15, or 20 depending on the course), many of the letter grades are impossible for students to earn on an individual assignment, and there is often only one possible midterm exam score for a student to earn each letter grade. For example, for midterm exams graded out of 15 points, it would be impossible for the student to earn a D+, B–, or B+ on that midterm exam based on the course grading scheme, and there is only a single score each that results in a D–, D, C–, C, C+, B, or A–. In addition, a student could move from a C to a C– or vice versa by answering one additional question correctly or incorrectly and presenting results that way may overemphasize transitions that are largely artifacts of our exam score structure. While our aggregated approach does not fully resolve it, many differences of a single question would remain in the same letter grade (e.g., A, B, C, D, or E).

For each midterm exam and its subsequent midterm exam (e.g., exam 2 and exam 3), we calculated the fraction of students who earned each letter grade on the subsequent midterm exam given the letter grade on the midterm exam of interest, essentially forming a “transition matrix” from one midterm exam grade to the next (similar to Morris *et al.* [31]). We then aggregated these fractions into the fraction of students who earned the same letter grade, earned a higher letter grade on the subsequent midterm exam, and earned a lower letter grade on the subsequent midterm exam. Students with missing midterm exam data were removed pairwise, meaning that a student could be missing a score on exam 3 but still be included in the exam 1 to exam 2 analysis.

Second, because the grade students earn on a midterm exam depends on both their performance and the underlying difficulty of the midterm exam, we also normalized each midterm exam score using the mean and standard deviation of each individual midterm exam to normalize. We grouped these normalized midterm exam scores into quintiles, picking five groups, as there are five potential letter grades. These quintiles allow us to compare how students perform relative to their peers across midterm

exams rather than only examine how their letter grades vary across midterm exams.

Finally, we repeated the analysis with the quintiles. We calculated the fraction of students in each quintile on the subsequent midterm exam given their quintile on the midterm exam of interest. We then aggregated these fractions into the fraction of students who fell in the same quintile, fell into a higher quintile on the subsequent midterm exam, and fell into a lower quintile on the subsequent midterm exam.

2. Alternative aggregation methods

To answer research questions 2 through 4, we conducted a thought experiment around how three alternative midterm exam score aggregation methods could affect student grades. As we needed all midterm exam grades (including the final as relevant), students with any missing midterm or final exam grades were excluded from this analysis. The number of students remaining for each assessment style is shown in Table I. Data on the students excluded from the analysis due to missing data are included in Section I in Supplemental Material [32].

In this paper, we examined three potential alternative aggregation methods based on our knowledge of what introductory physics instructors at other institutions have tried in their courses. These methods, how they were implemented and why instructors may use them are described below. It is important to note that none of these policies were actually implemented in the courses under study, and these represent various “what-if” scenarios based on our knowledge of what grading practices are used by physics instructors at other institutions.

- Drop the lowest midterm exam score: The student’s lowest midterm exam score is dropped, and the total weight of the midterm exams is split between the remaining midterm exams. For example, for a section with three midterm exams that in total contribute 36% to the final grade, the two best midterm exam scores would each contribute 18% of a student’s final grade. Such an approach allows a student to earn an abnormally low grade without lowering their overall grade in the course, whatever the reason (e.g., not understanding the course material or personal factors). However, in preliminary discussion of this manuscript with instructors, some worried that students would not study for or take the last midterm exam if they were already satisfied with their grades on the previous midterm exams because the last midterm exam score would be dropped regardless.
- Replace the lowest midterm exam score with the final: If a student scores better on the final exam than they did on their lowest midterm exam score, the final exam score replaces the lowest midterm exam score. Using the example in the previous bullet point, if the student scored higher on the final exam, the final

exam would take the place of the lowest midterm exam and count for 12% of the student's final grade (36%/3) *in addition* to the regular weighting of the final exam. Such an approach presumably rewards students for growth over time by allowing them to replace a low score with a higher score on the (cumulative) final exam.

- Weigh the best midterm exam score more than the lowest midterm exam score: The student's best score counts for 50% more and the student's lowest score counts for 50% less than the remaining midterm exams. For the same example section, the best midterm exam score would count for 18% of the final grade, the middle score would count for 12% of the grade, and the lowest score would count for 6% of the final grade. Such an approach combines the benefits of the previous two approaches, rewarding growth and limiting the impact of an abnormally low midterm exam score by weighting the highest midterm exam score the most and weighting the lowest midterm exam score the least. In addition, because all midterm exams still have a nonzero weight in the final course grade calculation, students cannot skip the last midterm exam without consequence, addressing an instructor concern about dropping the lowest midterm exam score.

The grading scheme used in the course (referred to here as *default*) divided the total weight of the midterm exams by the number of midterm exams to get the weight of each midterm exam. Again, using the example three midterm exam course worth 36% of the final grade, each midterm exam would have a weight of 12%.

For each student, we calculated their default aggregated midterm exam grade and their aggregated midterm exam grade under each modification. To calculate the default aggregated grade, all midterm exam scores as percentages used in the calculation were averaged. Because each midterm exam was worth the same number of points, this is equivalent to adding up the total points a student earned on midterm exams and dividing by the total possible points on the midterm exams.

To determine whether any of the three alternatives increased aggregated midterm exam scores relative to the default method, we conducted a t test between the aggregated midterm exam grade from the default midterm exam scoring method and the aggregated midterm exam grade from each of the three alternatives. To determine the practical significance of the differences, we also calculated the effect size using Cohen's d . We use the standard $\alpha = 0.05$ as our cutoff for statistical significance and $d \geq 0.2$ to be the minimum effect size for a non-negligible effect [35].

3. Final course grades and demographics

To answer research questions 3 and 4, we used final course grades and demographics from the university's data warehouse.

Grades are reported as letter grades with the plus-minus system. For the purposes of this paper, we grouped students who earned an A−, A, or A+ in Physics I overall as high-grade earners and all other students as lower-grade earners, similar to prior work [36]. We do this because when physics instructors refer to students who earn high grades in physics, they are generally referring to students who earn "A"s rather than, for example, students who earn "B"s even though those grades are still above average grades.

The university's data warehouse only collects data about sex and not gender, so results are split by sex. Sex is reported as male, female, or unknown in the data warehouse. For the purposes of this paper, we assume students who selected male are male and students who selected female are female, ignoring that sex is not binary [37,38]. We treat students marked as unknown as missing data. Our analysis then only compares students categorized as male with students categorized as female.

For race and ethnicity, students can select multiple options to describe their race and ethnicity. Students could select Asian, Black, Hispanic, Native American/Hawaiian, or white as possible options. We grouped Asian and white students into one group (called A/w) and all other students into a second group, following how similar studies have created binary racial groups [18,21]. Students who selected multiple racial or ethnic groups, regardless of which racial or ethnic groups were selected, were included in this second group. Because this group includes students who selected Black, Hispanic, Native American, and multiple categories, we called this group B/H/M/N. We acknowledge that aggregating groups of students can hide differences between groups [39].

As we are interested in differential effects, we computed the difference between the aggregated midterm exam score from each grading modification and the aggregated midterm exam score from the default grading scheme for each student, equivalent to the grade increase as a result of each modification. We then compared the average grade difference between each of the alternative midterm exam score aggregation methods and the default grading scheme for the two groups of students. To determine whether the differences were statistically significant, we performed t tests. We also calculated the effect size for each difference, again using Cohen's d . We did this for students divided by sex, then race, and then by final grade. For each analysis, we removed any student with missing sex, race, or final grade information only for that specific analysis. For example, a student who did not report their race but did report their sex and final grade would only be excluded from the race analysis. The number of students in each group for each analysis is shown in Table III.

In terms of our research framing, if we find that the grade difference between, for example, male and female

TABLE III. Number of students in each group and missing for each time period in the study.

Assessment styles	Male	Female	Sex missing	A/w	B/H/M/N	Race missing	Physics I grade: A	Physics I grade: B, C, D, E	Final grade missing
3E + F	1559	723	8	1780	374	136	599	1676	15
4E + F	791	479	4	940	257	77	371	894	9
4E	472	276	2	576	128	46	240	506	4
5E + F	513	251	3	587	127	53	162	603	2

students is not statistically different after implementing one of the alternative aggregation schemes, we would have evidence of equity of parity. If we do not find a statistically or practically significant difference between the grade increases by each of the alternative aggregation methods for each group, we would have evidence of equity of fairness.

4. Mixed-effects model

To account for the various course and demographic effects, we then ran a mixed-effects linear model. We first calculated the overall exam score difference between each of the three alternative exam aggregation methods and the equal-weighting method for each student. We refer to these differences as the *deltas*, Δ . We computed the following model:

$$\begin{aligned}\Delta_{ijk} = & \beta_0 + \beta_1 \text{Policy}_{ijk} + \beta_2 \mathbf{X}_i \\ & + \beta_3 (\text{Policy}_{ijk} \times \mathbf{X}_i) \\ & + u_i + v_j + \epsilon_{ijk},\end{aligned}$$

where **Policy** = (equal weighting, drop lowest, replace with final, and variable weighting) and $\mathbf{X}$ = (sex, race, and final grade), with each of these as a binary variable, u_i is a student-level random intercept, v_j is the assessment-style random intercept, and ϵ_{ijk} is the residual error. We included the student-level and assessment-style random intercepts to account for differences between individual students and differences between the number of exams in the course.

We included equal weighting as a policy with $\Delta = 0$ so that model coefficients for the remaining policies represent the change in average exam grade relative to the equal weighting. We take male and A/w as the baseline levels for the demographic variables and a final grade of “A” as the baseline for the final grade.

In this model, the β_1 values address RQ2, and the β_3 values address RQs 3 and 4. If we find the β_1 ’s statistically different than zero, we have evidence that the alternative exam aggregation methods increase average exam grades. If we find the β_3 ’s statistically different than zero, we have evidence that there are differential effects based on either sex, race, or final grade in the course.

It is important to note that the β_1 ’s are relative to the reference groups. Therefore, to find the change in aggregated exam grades from equal weighting for students not in the reference groups, we would add the coefficient for the relevant interactions.

5. Multiple comparisons

Throughout this paper, we report the results of multiple hypothesis tests, and as such, we need to consider multiple comparison issues and adjust p values. Following the recommendation of Boulesteix and Hoffmann [40], we do not use any adjustment procedures, as we report the results of all performed tests regardless of their outcome and we do not weight a single result more than others.

IV. RESULTS

A. Grade variations

Overall, we find 43% of students earn a lower letter grade on their subsequent midterm exam, 22% earn a higher letter grade on their subsequent midterm exam, and 35% earn the same letter grade on their subsequent midterm exam.

Breaking the results out by each midterm exam, Fig. 1(a), we find that in a majority of the assessment styles and midterm exam transitions, the most common outcome is still that students earn a lower grade on the subsequent midterm exam. This result occurred in 7 of the 12 midterm exam transitions. Of the remaining five midterm exam transitions, the most common result was students earning the same grade on the subsequent midterm exam (two cases), followed by earning a higher grade on the subsequent exam, earning the same grade or a higher grade, and earning the same grade or a lower grade in one case each.

Examining quintiles overall, we find that 35% of students place in the same quintile on the subsequent midterm exam, 33% place in a higher quintile on the subsequent midterm exam, and 32% place in a lower quintile on the subsequent midterm exam.

Breaking the results out by each midterm exam again, Fig. 1(b), we still find that remaining in the same quintile on the subsequent midterm exam is the most common result, occurring in 6 of the 12 transitions. We find that having approximately the same percentage of students move up a quintile as moving down a quintile is the second

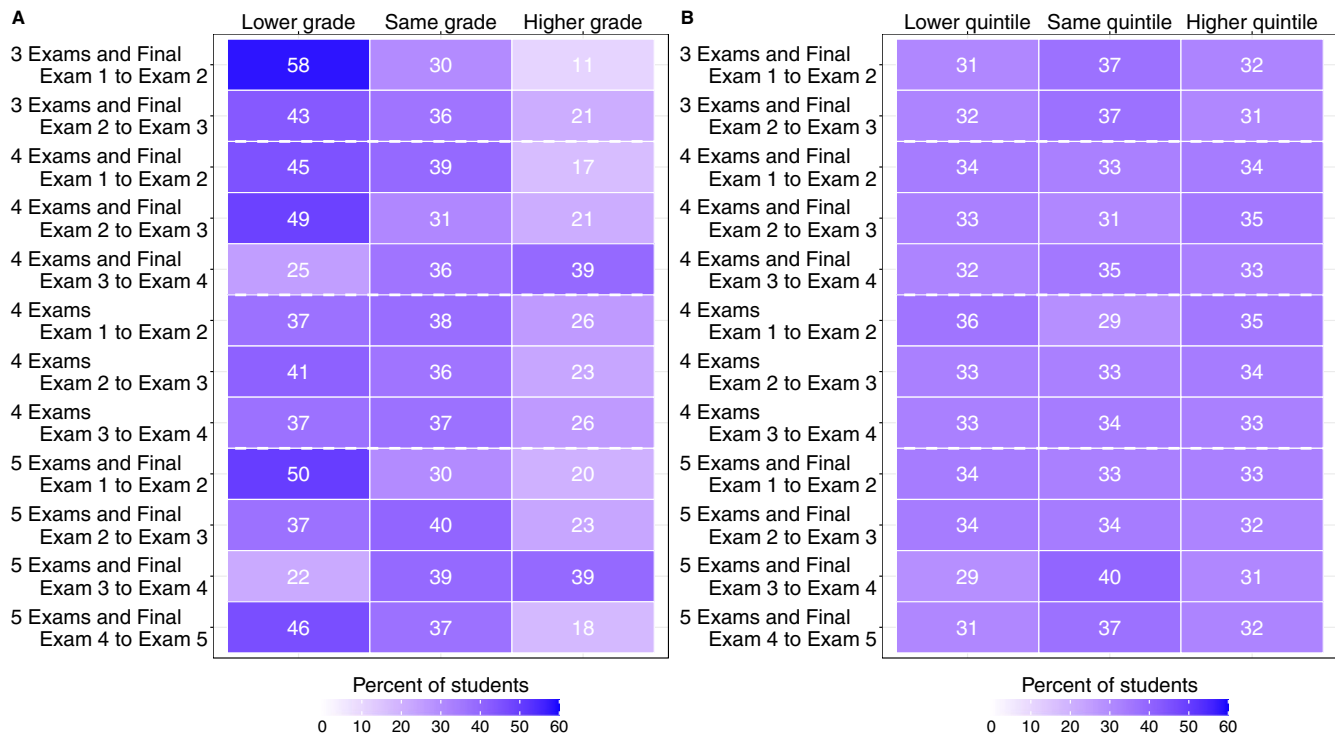


FIG. 1. Percent of students earning a lower letter grade, the same letter grade, or a higher letter grade on each subsequent midterm exam (plot A) and students in a lower quintile, the same quintile, and a higher quintile on each subsequent midterm exam (plot B). Numbers in the boxes are the percentage of students. Rows may not add up to exactly 100% due to rounding. Students earning a lower letter grade on the subsequent midterm exam was the most common grade transition, and students being in the same quintile on the subsequent midterm exam was the most common quintile transition.

most common result, occurring in 3 of the 12 transitions. We also found a more uniform distribution of outcomes when examining quintiles compared to the letter grades, with each transition having at least 29% of the students and no more than 40% of the students for the quintiles, while at least one of the possible outcomes (lower, same, or higher) had fewer than 29% of students for all 12 transitions when looking at letter grades.

Letter grade-specific transitions are included in Section IV in Supplemental Material [32].

B. Alternative aggregation methods

Results from the alternative midterm exam aggregation methods are shown in Fig. 2. We find that dropping the lowest midterm exam score resulted in the largest increase in grades under all four assessment styles, with a minimum increase of 4.5% points. In course sections with a final exam, we found that replacing the lowest midterm exam score with the final exam score if higher and variably weighting the midterm exam scores resulted in a similar increase in grades.

We find that each of the aggregated grades is statistically different from the default aggregated midterm exam grade, as the 95% confidence intervals do not overlap. We also find that implementing each of the alternative aggregation

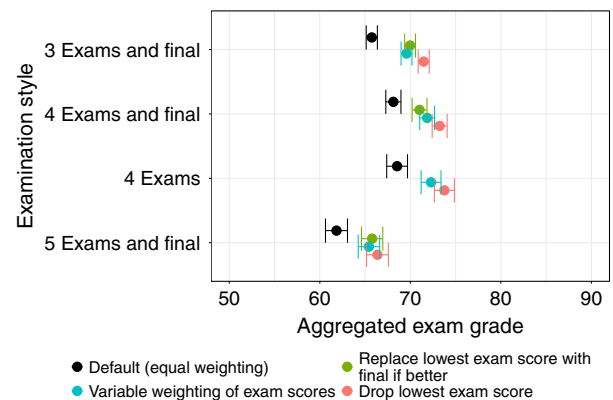


FIG. 2. Average aggregated midterm exam grade for the default equal weighting and each of the three alternative methods for the four assessment styles. Error bars represent a 95% confidence interval. Raw exam scores were used to calculate the aggregated exam grades. All alternative methods increase grades, with dropping the lowest midterm exam score, resulting in the largest increase in grades.

methods would have a small effect size (> 0.2), with the exception of replacing the lowest midterm exam with the final in the four midterm exams and a final period, which had no effect. These effect sizes are shown in Table IV.

TABLE IV. Effect sizes (Cohen's d) of the three alternative midterm exam aggregation methods compared to the default equal weighting. Positive values mean that the alternative midterm exam aggregation method resulted in higher midterm exam scores than the equal-weighting method. Numbers in parentheses represent the 95% confidence interval.

Assessment style	Drop lowest vs default	Replace with final vs default	Variable weighting vs default
Three midterm exams and final	0.38 (0.31, 0.43)	0.27 (0.21, 0.33)	0.26 (0.20, 0.32)
Four midterm exams and final	0.34 (0.26, 0.42)	0.19 (0.11, 0.27)	0.24 (0.17, 0.33)
Four midterm exams	0.33 (0.23, 0.43)	NA	0.24 (0.14, 0.34)
Five midterm exams and final	0.27 (0.17, 0.37)	0.23 (0.13, 0.33)	0.21 (0.11, 0.31)

C. Final grades

When examining the results based on the final grade earned in Physics I (Fig. 3), we still find that dropping the lowest midterm exam scores results in the largest aggregated midterm exam score increase and that the students who earned “A”s in Physics I always outperform the other students on average.

Examining the differential impacts, we now find that students who earned a Physics I grade of “B” or lower

differentially benefited from these alternative aggregation methods. We find a statistically significant difference of at least 0.45% points and as much as 2.13% points in 10 of the 11 aggregation-method-and-assessment-style pairs. Further, 9 of these 10 aggregation-method-and-assessment-style pairs also had at least a small effect size (> 0.2), and 4 of them had a large effect size (> 0.8). The aggregation-method-and-assessment-style pairs that were not statistically significant or did not have at least a small

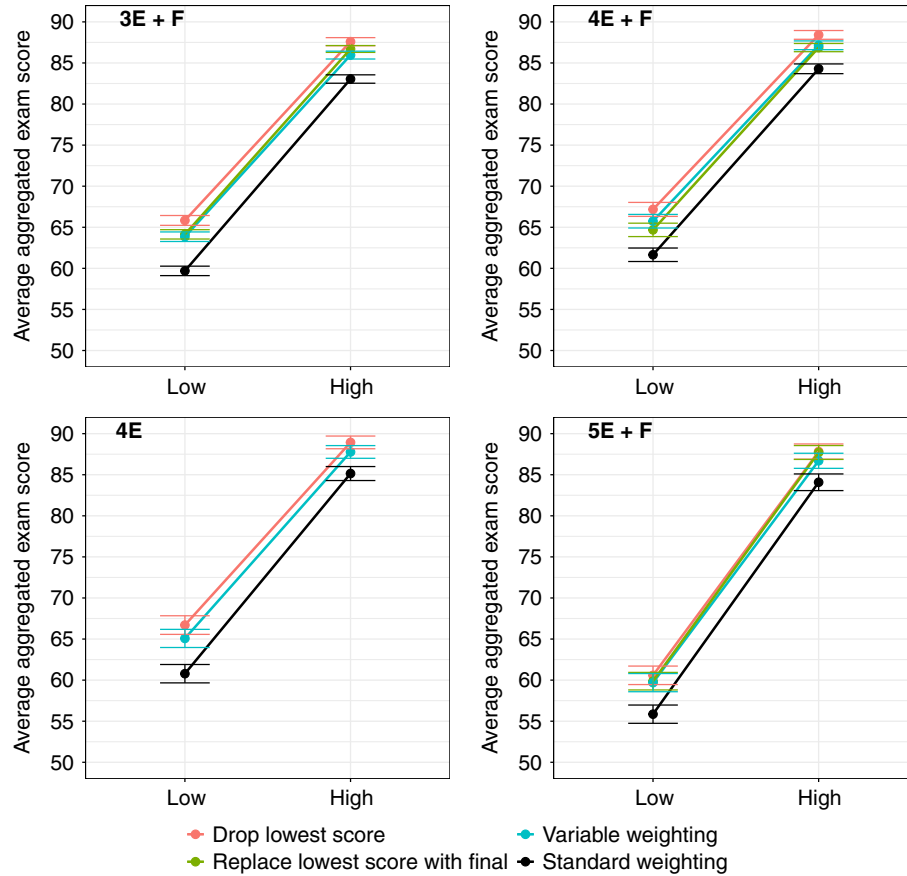


FIG. 3. Average aggregated midterm exam scores for students who earned an “A” in Physics I (high) and students who earned a “B,” “C,” “D,” or “E” in Physics I (low) according to the aggregation method used and the assessment style. Note that the vertical axis does not start at zero. Error bars represent a 95% confidence interval. Raw exam scores were used to calculate the aggregated exam grades. In all courses, “A” students outperformed the other students, but students who earned a “B,” “C,” “D,” or “E” generally received a larger benefit.

TABLE V. Difference in increases in aggregated midterm exam scores compared to the default method between students who earned an “A” in Physics I and students who earned a “B,” “C,” “D,” or “E” in Physics I (A), and effect sizes (B). Positive values mean that students who earned an “A” received a larger benefit, and the grade gap increases. The values represent the aggregated midterm exam score differences (reported in percentage points), with $p < 0.001$ represented as ***, $p < 0.01$ as **, and $p < 0.05$ as * and 95% confidence intervals in parentheses. Values in (B) represent the effect size (Cohen’s d), nearly all of which are non-negligible.

Format	Drop lowest	Replace with final	Variable weighting
(A) Grade differences			
Three midterm exams and final	−1.59*** (−1.86, −1.33)	−0.79*** (−1.11, −0.46)	−1.24*** (−1.41, −1.08)
Four midterm exams and final	−1.40*** (−1.67, −1.12)	−0.45** (−0.78, −0.12)	−1.23*** (−1.40, −1.06)
Four midterm exams	−2.13*** (−2.53, −1.73)	NA NA	−1.66** (−1.90, −1.43)
Five midterm exams and final	−0.99*** (−1.36, −0.62)	−0.39 (−0.87, −0.08)	−1.22*** (−1.45, −1.00)
(B) Effect sizes			
Three midterm exams and final	−0.53 (−0.62, −0.43)	−0.21 (−0.30, −0.13)	−0.66 (−0.75, −0.57)
Four midterm exams and final	−0.59 (−0.70, −0.46)	−0.16 (−0.28, −0.04)	−0.82 (−0.94, −0.71)
Four midterm exams	−0.80 (−0.98, −0.64)	NA NA	−1.05 (−1.22, −0.88)
Five midterm exams and final	−0.47 (−0.66, −0.28)	−0.14 (−0.31, 0.04)	−0.89 (−1.07, 0.70)

effect size were found under the replace the lowest midterm exam score with the final. These are shown in Table V.

D. Demographics

Starting by comparing male and female students, we find similar results to the last section in that dropping the lowest midterm exam score continues to result in the largest grade increase (Fig. 4).

Across all assessment styles, we find that on average, male students earned statistically higher midterm exam grades than female students. We find this result to be true regardless of whether we use the default equal-weighting aggregation scheme or whether we use one of the three alternative aggregation schemes. In fact, the average scores of male students under the standard weighting were similar to the average scores of female students under one of the alternative aggregation schemes.

The differences in increases in grades between male and female students are shown in Table VI(A). We find these differences to be relatively small, with the largest difference being about half a percentage point. We find that 8 of the 11 aggregation-method-and-assessment-style pairs do not show a statistically significant difference in the increase in grades between male and female students. We find that 2 of the 11 aggregation-method-and-assessment-style pairs show a benefit to male students and 1 of the 11 aggregation-method-and-assessment-style pairs shows a

benefit to female students. We find all of the effect sizes to be less than 0.2, resulting in a negligible effect size. These are shown in Table VI(B).

Examining the data split by race, we find similar results (Fig. 4). We still find that dropping the lowest midterm exam score results in the largest grade increase. For all assessment styles and for all alternative aggregated methods, we find that A/w students outperform B/H/M/N students. We also find that the average scores of A/w students under the standard weighting were similar to the average scores of B/H/M/N students under one of the alternative aggregation schemes.

The differential impacts of the alternative midterm exam aggregation methods are shown in Table VII(A). We find that none of the differential impacts are statistically significant, meaning that there are no statistical differences between Asian and white students and Black, Hispanic, multiracial, and Native American students. The effect sizes, also shown in Table VII(B), are negligible, further suggesting minimal differential impacts.

E. Mixed-effects model

Our results are shown in Table VIII.

First, we find that all of the coefficients for the alternative aggregation methods are positive and statistically different from zero. They range from 2.877 to 4.294, meaning that aggregated exam grades increase by 2.877 to 4.294 points compared to the equal weighting

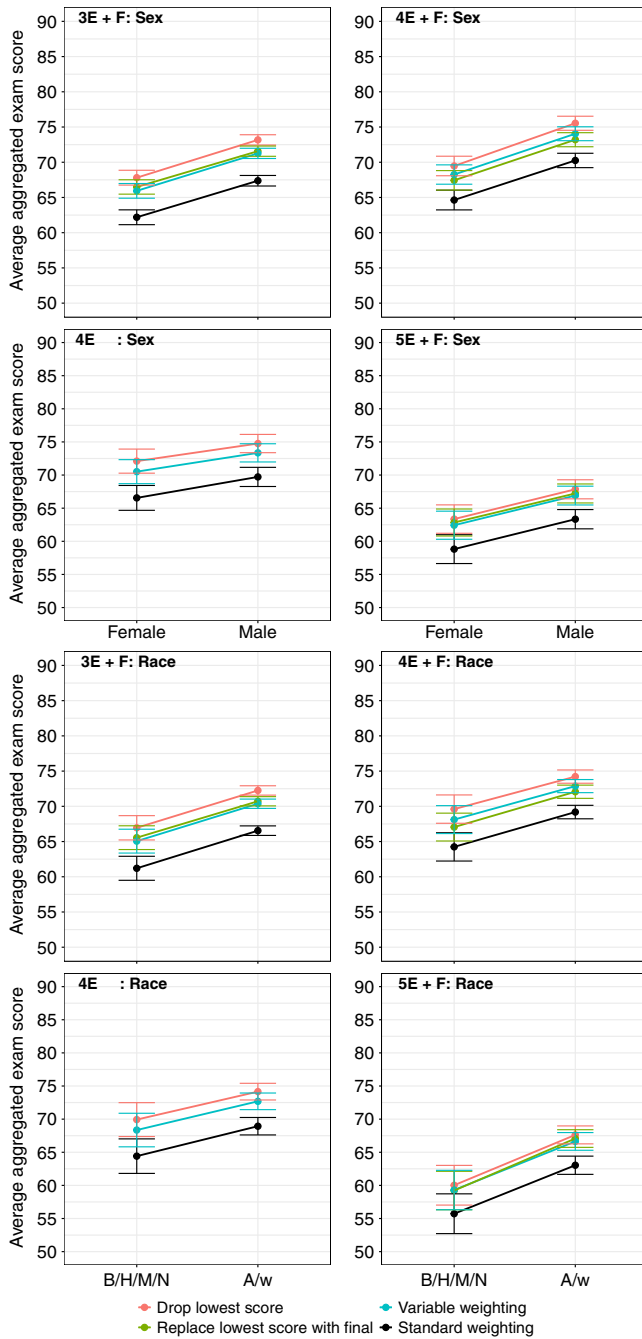


FIG. 4. Average aggregated midterm exam scores by sex and race according to the aggregation method used and the assessment style. Note that the vertical axis does not start at zero. Error bars represent a 95% confidence interval. Raw exam scores were used to calculate the aggregated exam grades. In all course sections, male students outperformed female students, and Asian and white students outperformed Black, Hispanic, multiracial, and Native American students. We found limited evidence of differential effects based on sex or race, however.

used in the course for Asian and white male students who earned a final grade of “A” in the course. Similar to our previous results, we find that dropping the lowest exam score results in the largest increase. However, now

replacing the lowest exam score with the final exam score clearly results in a larger aggregated exam score increase than variably weighting the exams.

Second, we find the coefficients of sex, race, and final grade are not statistically different than zero. We expect this to be the case because of our model construction. When none of the alternative exam aggregation policies are in effect, the outcome variable, Δ , should be equal to zero for all students because the exam grade for the student has not changed from the baseline value. Therefore, all of these coefficients should be approximately zero.

Third, we examine the interactions between the alternative exam aggregation policies and sex. We find evidence that the interactions between dropping the lowest exam grade and variably weighting the exams and sex are negative and statistically significant. These results suggest that female students may receive a smaller benefit than male students do from these alternative exam aggregation policies. However, the coefficients of these interactions are less than 6% the size of alternative exam aggregation policy coefficients, suggesting that inequities introduced are relatively small compared to the overall benefit of the alternative exam aggregation policies on increasing grades.

Fourth, we examine the interactions between the alternative exam aggregation policies and race. We do not find evidence that the coefficients of any of these interactions are statistically different from zero, as zero is contained in all of the 95% confidence intervals.

Finally, we examine the interactions between the alternative exam aggregation policies and final grades. We find that all the interaction coefficients are positive and statistically significant. These results suggest that students who did not earn a final grade of “A” in the course receive a larger benefit compared to students who earned an “A” in the course. The coefficients range from approximately 20% to approximately 45% of the overall benefit of the alternative exam aggregation policies on increasing grades.

V. DISCUSSION

Here, we provide answers to our research questions and discuss them individually.

A. How do student grades vary across midterm exams within an introductory physics course?

We find that student grades tend to vary across midterm exams within an introductory physics course. Overall, the most common midterm exam grade transition for students was to earn a lower grade on their subsequent midterm exam, while the most common quintile transition was for students to remain in the same quintile on the subsequent midterm exam. However, the majority of students moved to a higher or lower quintile, with about equal fractions of

TABLE VI. Difference in increases in aggregated midterm exam scores compared to the default method between male and female students (A) and the effect size (B). Positive values mean that males received a larger benefit, and the grade gap increases in both tables. Values in (A) represent aggregated midterm exam score differences (reported in percentage points), with $p < 0.001$ represented as ***, $p < 0.01$ as **, and $p < 0.05$ as * and 95% confidence intervals in parentheses. Values in (B) represent the effect size (Cohen's d), all of which are negligible.

Format	Drop lowest	Replace with final	Variable weighting
(A) Sex differences			
Three midterm exams and final	-0.18 (-0.46, 0.10)	0.11 (-0.24, 0.45)	-0.14 (-0.32, 0.40)
Four midterm exams and final	-0.45** (-0.72, -0.17)	-0.15 (-0.46, 0.18)	-0.17 (-0.35, 0.01)
Four midterm exams	0.51* (0.08, 0.95)	NA NA	0.33* (0.06, 0.60)
Five midterm exams and final	0.02 (-0.30, 0.35)	0.13 (-0.32, 0.59)	0.04 (-0.19, 0.27)
(B) Effect sizes			
Three midterm exams and final	-0.06 (-0.14, 0.03)	0.03 (-0.06, 0.12)	-0.07 (-0.16, 0.02)
Four midterm exams and final	-0.18 (-0.29, -0.07)	-0.05 (-0.17, 0.06)	-0.10 (-0.22, 0.01)
Four midterm exams	0.18 (0.03, 0.35)	NA NA	0.18 (0.04, 0.32)
Five midterm exams and final	0.01 (-0.15, 0.16)	0.04 (-0.10, 0.19)	0.03 (-0.12, 0.18)

TABLE VII. Difference in increases in aggregated midterm exam scores compared to the default method between Asian and white students and Black, Hispanic, multiracial, and Native American students (A) and the effect size (B). Positive values mean that Asian and white students received a larger benefit and the grade gap increases. Values in (A) represent aggregated midterm exam score differences (reported in percentage points), with $p < 0.001$ represented as ***, $p < 0.01$ as **, and $p < 0.05$ as * and 95% confidence intervals in parentheses. Values in (B) represent the effect size (Cohen's d), all of which are negligible.

Format	Drop lowest	Replace with final	Variable weighting
(A) Race			
Three midterm exams and final	-0.04 (-0.42, 0.34)	-0.18 (-0.61, 0.26)	-0.03 (-0.27, 0.20)
Four midterm exams and final	-0.33 (-0.69, 0.03)	0.07 (-0.34, 0.48)	-0.19 (-0.42, 0.04)
Four midterm exams	-0.30 (-0.86, 0.27)	NA NA	-0.18 (-0.53, 0.17)
Five midterm exams and final	0.28 (-0.12, 0.68)	0.53 (-0.04, 1.10)	0.02 (-0.27, 0.31)
(B) Effect sizes			
Three midterm exams and final	-0.01 (-0.13, 0.11)	-0.04 (-0.16, 0.07)	-0.02 (-0.12, 0.10)
Four midterm exams and final	-0.13 (-0.27, 0.01)	0.03 (-0.11, 0.16)	-0.11 (-0.25, 0.02)
Four midterm exams	-0.10 (0.30, 0.08)	NA NA	-0.10 (-0.29, 0.10)
Five midterm exams and final	0.13 (-0.05, 0.31)	0.18 (-0.01, 0.38)	0.01 (-0.20, 0.21)

students moving in each direction. The distribution of transitions was also much more uniform, with the most common transition overall, staying in the same quintile, occurring only 3% points more than the least common

overall, moving to a lower quintile. These results largely align with Suresh and Heckler's finding of significant within-student changes in relative grade standing between midterm exams [21].

TABLE VIII. Coefficients, standard deviation, lower and upper limits of the 95% confidence interval, and p value for our mixed-effect model, with $p < 0.001$ represented as ***, $p < 0.01$ as **, and $p < 0.05$ as *.

Variable	Estimate	SD	Lower CI	Upper CI	p value	
Intercept	-0.074	0.134	-0.352	0.202	0.604	
Drop lowest	4.294	0.080	4.139	4.450	< 0.001	***
Replace with final	3.299	0.084	3.134	3.464	< 0.001	***
Variable weight	2.877	0.080	2.721	3.033	< 0.001	***
Sex (male = 0)	0.017	0.072	-0.125	0.159	0.814	
Race (A/w = 0)	0.022	0.088	-0.152	0.194	0.808	
Final grade (A = 0)	0.007	0.077	-0.145	0.158	0.932	
Drop lowest $\times$ sex	-0.248	0.083	-0.410	-0.086	0.003	**
Replace with final $\times$ sex	-0.046	0.087	-0.217	0.125	0.600	
Variable weight $\times$ sex	-0.164	0.083	-0.326	-0.002	0.048	*
Drop lowest $\times$ race	-0.071	0.101	-0.269	0.127	0.481	
Replace with final $\times$ race	-0.119	0.106	-0.326	0.089	0.263	
Variable weight $\times$ race	-0.039	0.101	-0.237	0.158	0.696	
Drop lowest $\times$ final grade	1.528	0.089	1.355	1.702	< 0.001	***
Replace with final $\times$ final grade	0.700	0.094	0.516	0.884	< 0.001	***
Variable weight $\times$ final grade	1.292	0.089	1.118	1.465	< 0.001	***

Of the students who did not earn the same letter grade on the subsequent midterm exam, we found that the most common result was to change by one grade step compared to the previous midterm exam (see Section IV in Supplemental Material [32]). Such results largely align with other studies that examine how student grades on high-stakes exams evolve over a semester [41,42].

Overall, our results suggest that while a student's grades across midterm exams are related, they are not necessarily deterministic: students who earn lower grades early on can earn a higher grade in the next midterm exam and vice versa. However, we note that large improvements in grades are relatively rare and that most students who perform poorly on the first midterm exam do not become higher performers on the next midterm exam. From an instructional standpoint, this suggests that most students will not drastically improve their midterm exam grades on their own. As these are actual grades that already take into account existing resources offered by the instructor, supporting students to drastically improve their midterm exam grades will require resources and strategies beyond those already offered in the course.

It is important to note that these results could be explained due to exams becoming more challenging as the semester progresses or the content covered becoming more challenging for students rather than attributed to student ability. Average exam scores do, in general, decrease (see Section II in Supplemental Material [32]), and this suggests that students earning a lower grade may be more about the difficulty of the exam rather than student ability. This interpretation is further supported by the most common exam transition being students remaining in the same quintile, meaning that their performance relative to their peers is similar between exams. Therefore, the apparent decline is likely driven by exam difficulty.

B. What impact do alternative aggregation methods have on students' overall midterm exam grades as compared to weighting each midterm exam equally?

Unsurprisingly, we find that all three alternative midterm exam aggregation methods we examined increased student midterm exam scores relative to the equal-weight aggregation method used in each course section of our study.

Using descriptive statistics, we found that dropping the lowest midterm exam score resulted in the largest grade increase, while replacing the lowest midterm exam score with the final exam score (if higher) or weighting each midterm exam differently resulted in similar grade increases to each other. In terms of magnitude, the minimum grade increase we observed was 3.5% points, which occurred when replacing the lowest midterm exam score with the final in the five midterm exams and a final format condition, and the largest grade increase we observed was 5.5% points, which occurred when dropping the lowest midterm exam in the three midterm exams and a final condition. Across the four assessment styles, we generally found the effect size of implementing these alternative midterm exam aggregation methods to be small.

Our mixed model largely found the same results. We found that all three methods did increase students' aggregated exam grade, with dropping the lowest midterm score having the largest effect, followed by replacing the lowest midterm with the final exam score if higher and finally, variably weighting the exams.

To contextualize these percentage increases, we can examine the impact on students' final grades. As the midterm exams in these Physics I courses tended to account for 36% of the student's final grade in the course, we estimate

that implementing these practices could raise final grades by between 1.25 and 1.98 points (using the descriptive statistics estimates since classroom grades do not control for other factors), potentially enough to raise students a letter grade (for example, a 73.5 C+ could be raised to a B– using the grading scale of this course). For courses where midterm exams constitute a larger percentage of the final grade, the effect would be even greater. In our course, we estimate that between 20% and 40% of students in a course would experience such a letter grade jump.

While such a grade increase may seem trivial, there can be potential real-world consequences. For example, a student earning a B– instead of a C+ would earn an additional 0.4 grade points, which could potentially matter for students on the edge of scholarship eligibility or maintaining grade point average minimums. Further, recent work suggests students who score just above a grade cutoff in an introductory course are slightly more likely to major in that subject than students who do not score just above a grade cutoff [43], suggesting that small changes in grading policy could influence enrollment and decisions around a student’s major.

Of course, these results rest on the assumption that student behavior would be unchanged if any of these alternative midterm exam aggregation methods were implemented in the course. If student behavior did change, then the grade increases could be different. For example, if students know their lowest midterm exam grade is dropped, they may not study for the last midterm exam if they are already happy with their aggregated grade, resulting in a lower grade increase than found here and potentially less learning, a concern frequently raised by colleagues in preliminary discussions of the results of this paper.

While we cannot directly address this concern with our data, prior work outside of physics suggests that this may not happen in practice. For example, one study found that the majority of students do not engage in any sort of “strategic behavior” when midterm exam scores are dropped [44], and another found that telling students their lowest midterm exam score would be dropped did not influence their decisions to miss or not try their best on a midterm exam [45]. The effect of dropping midterm exam grades on final exam scores, assumed to be a measure of overall learning in the course, is less clear, as some studies have found students scored better on final exams when midterm exam scores were dropped [46,47], while other studies have found lower final exam grades [45].

Alternatively, if students are aware that improved performance on a later midterm exam or the final exam could mitigate the impact of an earlier bad grade, perhaps students would work harder than they otherwise would. Evidence in support of this hypothesis is mixed, as some studies have found that students near grade boundaries perform better than students not near grade

boundaries [48,49], while others have found no difference [50]. Other studies suggest that students who have low grades near the middle of the term do not focus on studying for final exams as a way to improve their final grade but rather focus on the lower-stakes aspects of the course, such as participation and in-class activities [51]. Whether this would happen in practice in physics courses and courses with multiple midterms, such as the course studied here, is again an open question.

Finally, while we have shown that the three alternative midterm exam aggregation methods increase average midterm exam scores, we make no recommendation as to whether increasing grades is necessarily an ideal outcome or should be the ideal outcome. Debates around grade inflation in postsecondary education are hardly new [52–55], as are debates about curving exam grades [56–58]. Our goal for this paper is not to necessarily argue that all students should receive higher grades in their introductory physics courses but rather to call attention to how course management decisions unrelated to student learning can have a potentially meaningful impact on the grades students earn.

C. How do the impacts of the alternative aggregation methods vary in terms of the students’ final grade?

We found that both students who earned “A”s and “B”s or lower in Physics I could benefit from alternative midterm exam aggregation policies.

Using descriptive statistics, we found that in most of the course assessment styles we examined, students who earned a Physics I grade of “B” or lower would differentially benefit from these policies. Upon dropping the lowest midterm exam grade or variably weighting the midterm exam grades, students who earned a “B” or lower received an additional grade bump of at least 1% point compared to students who did earn “A”s.

Our mixed model also finds similar effects. All the interactions between alternative exam aggregation policies and final grade were positive and statistically significant. The size of the interactions ranged from 0.700% to 1.528% points, reflecting a similar magnitude of the impact as the descriptive statistics analysis.

As these alternative exam aggregation policies are often intended to help students who may be struggling in the course, this is perhaps not surprising. Similar results have been observed before, including one study that found students with the highest and lowest scores did not significantly benefit from grade dropping [47]. While students who earned a “B” or lower would benefit more from any of the alternative midterm exam aggregation methods than students who did earn “A”s, students who earn “A”s still significantly outperform the other students, suggesting that implementing any of these methods would not hamper an instructor’s ability to identify top-performing students.

D. How do the impacts of the alternative aggregation methods vary based on the students' sex or race?

In terms of race, we find no evidence that alternative exam aggregation policies benefited Black, Hispanic, multiracial, or Native American students more than Asian and white students or vice versa. Both the descriptive statistics and mixed-effects model are consistent with the hypothesis that both groups of students are impacted equally by these policies. As a result, the alternative midterm exam aggregation methods explored here are equitable under the equity of fairness approach, as both racial groups receive approximately the same benefit. However, as all groups receive approximately the same benefit, preexisting equity gaps will be largely unaffected. Because preexisting equity gaps are unaffected, we do not find evidence of equity of parity.

The interpretation of sex is a bit more complicated. From the descriptive statistics analysis, we did not find a consistent effect where any of these methods benefited female students more than male students or vice versa. From the mixed-effects model, we do find evidence that these alternative exam aggregation policies may slightly benefit male students more than female students. However, the strength of this difference is relatively small in comparison to the main effects of the alternative exam aggregation policies. For example, male students receive a grade increase of 4.294% points on average by dropping the lowest exam score. Female students receive a grade increase of 4.046% points ($4.294 + -0.248$) or $\sim 94\%$ of the benefit that male students receive. As the midterm exams accounted for 36% of the final grade, this 0.248% point difference in aggregated exam scores would result in a 0.09 point difference in final grades on average between male and female students, which is practically negligible. For instructors especially concerned about the practical impacts of this, we recommend manually examining the final grades of students near a grade cutoff to determine if their grade should be rounded up.

Taking these results together, we conclude that while there may be a small differential effect, our results are largely consistent with an equity of fairness conclusion in that both groups benefit approximately equally for all practical purposes. As all groups receive approximately the same benefit, preexisting equity gaps will still be largely unaffected. Because preexisting equity gaps are unaffected, we still do not find evidence of equity of parity.

If instructors wanted to achieve equity of parity, it would require rethinking the structure of exams. For example, recent work suggests that allowing exam retakes or reducing the weight of exams in the final grade could eliminate gender equity gaps on exams [20]. Other work suggests that addressing test anxiety could make exams more equitable [59], as could focusing on conceptual understanding before performing complicated mathematical calculations [19].

VI. LIMITATIONS AND FUTURE WORK

Our study has multiple limitations that might affect the generalizability and applicability of our results. We discuss them here as well as future work that could resolve some of these concerns.

First, this study was a thought experiment that assumed student behavior would be unchanged if these alternative midterm exam aggregation methods were implemented in class. If students did change their behavior in response, the effects reported here could be larger or smaller depending on whether students became more or less motivated to perform well on their midterm exams. In addition, if an exam had no impact on the final grade (because it is dropped or replaced with the final), students may choose to take a low grade or zero on the exam. In that case, our assumption about not changing student behavior may not be accurate.

Future work could address this limitation by collecting midterm exam grades from courses where these alternative midterm exam aggregation methods are implemented, effectively accounting for any changes to student behavior. By recalculating the grades without the alternative midterm exam aggregation method, we could obtain an estimate of how much these methods impact student grades.

Alternatively, for standardized courses with common midterm exams, we could compare aggregated student midterm exam grades in sections that implemented an alternative midterm exam aggregation method to sections that did not. Such an approach would also allow us to account for student behavior.

Second, the data for this study came from a subset of students taking Physics I rather than all students in the course. As a result, this study may not accurately represent the students who took Physics I if the students who are choosing to enroll in ECoach are fundamentally different from those in the course. As students choose to enroll in ECoach early in the semester before taking any midterm exams, we would not expect earned grades to influence enrollment in ECoach. However, students who expect to do exceptionally well in the course may choose not to use ECoach because they do not believe they need the extra help, resulting in students with higher grades potentially missing from the data. More than 90% of Physics I students in sections that offered ECoach did enroll in ECoach and were included in this study, suggesting that any impact of missing students would likely be minimal.

To account for this, future work could collect midterm exam data directly from the gradebook so that all students can be included in the data collection. We note that whether this is possible will likely depend on institutional policies around waiving informed consent on accessing student records. Otherwise, if informed consent is required, researchers may encounter the same limitation as here and be unable to obtain a 100% consent rate.

Finally, this study occurred at a single institution, and the results may not be generalizable to other universities with different student populations or with different grading scales in Physics I. For example, the demographics of the students included in this study are not representative of the physics-taking population as a whole [60]. In addition, while the midterm exams are still considered “high-stakes” at this university, the midterm exams and the final exam only constitute slightly more than half of the student’s final grade in the courses studied here. For courses that have a significantly higher fraction of final grade determined by midterm exam scores (e.g., the course studied in Simmons and Heckler, where the exam components constituted 70% of the final grade [18]), students’ behaviors may also change in ways different from how they might change if the lowest grade were dropped, for example. That is, if midterm exams make up a significant portion of their final grade, students may study more for the midterm exams than if they only make up about half of the final grade or even less.

Future work could conduct studies similar to this one at a variety of institutions and in a variety of physics departments to understand how local influences may affect the results. While we have examined four different examination styles here, we have left the weight of the aggregated midterm exam score in the final grade largely untouched.

More generally, future work could also expand beyond the grades themselves to focus on student behavior. For example, future work could interview students in courses implementing alternative midterm exam aggregation methods to understand how student behavior might be affected. For example, alternative midterm exam aggregation methods might impact student motivation, stress, and exam anxiety, the last of which is known to have a gendered impact on exam performance in introductory physics [61].

Further, for students who do significantly improve their midterm exam grades over time, future work could examine the strategies those students are using in comparison to strategies employed by students whose grades do not improve over time. Results from such a study could be helpful for developing interventions for students in danger of failing a course and could potentially be implemented in a platform such as ECoach, allowing the interventions to be delivered at scale to students.

VII. CONCLUSION AND IMPLICATIONS

In this paper, we have shown that student midterm exam grades in Physics I exhibit variation over time, with students typically increasing or decreasing their midterm

exam grades by at least one letter grade. We then showed that alternative midterm exam aggregation methods such as dropping the lowest midterm exam score, replacing the lowest midterm exam score with the final exam score if the student performed better, and weighting the highest midterm exam score more than the lowest midterm exam score in the final grade are all effective methods for increasing the average aggregated midterm exam grade in a course, with dropping the lowest midterm exam score resulting in the largest midterm exam increase. Depending on the weight of midterm exams in the final grade, these modifications could increase students’ final course grades by almost 2% points.

We also examined these alternative midterm exam aggregation methods as potential methods to address equity gaps. We find no evidence to support that these methods could reduce the size of equity gaps. We find no evidence in support of a differential benefit to Asian and white students compared to Black, Hispanic, multiracial, and Native American students or vice versa. We find limited evidence that these alternative exam aggregation policies may benefit male students more than female students, though the effect is relatively small in comparison to the overall grade increase from alternative exam aggregation policies.

Taken together, this work finds that the alternative midterm exam aggregation methods explored here can be effective for raising average course grades, and students who earn lower grades especially benefit. However, if an instructor is hoping to reduce equity gaps in their course, we do not recommend implementing the alternative midterm exam aggregation methods explored here and instead suggest instructors explore alternative structural changes to their courses. At the same time, any harm from implementing the alternative midterm exam aggregation methods explored here seems minimal.

ACKNOWLEDGMENTS

We thank Erin Murray for her assistance in securing the data for this project. Eric F. Bell was partly supported by the National Science Foundation through Grant No. NSF-AST 2007065. No additional external or internal funding was received for this study. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

DATA AVAILABILITY

The data that support the findings of this article are not publicly available. The data are available from the authors upon reasonable request.

- [1] *Talking About Leaving Revisited: Persistence, Relocation, and Loss in Undergraduate STEM Education*, edited by E. Seymour and A.-B. Hunter (Springer International Publishing, Switzerland, 2019).
- [2] B. Ost, The role of peers and grades in determining major persistence in the sciences, *Econ. Educ. Rev.* **29**, 923 (2010).
- [3] K. Rask, Attrition in STEM fields at a liberal arts college: The importance of grades and pre-collegiate preferences, *Econ. Educ. Rev.* **29**, 892 (2010).
- [4] R. Stinebrickner and T.R. Stinebrickner, A major in science? Initial beliefs and final outcomes for college major and dropout, *Rev. Econ. Stud.* **81**, 426 (2014).
- [5] S. D. Castle *et al.*, Systemic advantage has a meaningful relationship with grade outcomes in students' early STEM courses at six research universities, *Int. J. STEM Educ.* **11**, 14 (2024).
- [6] R. L. Matz, B. P. Koester, S. Fiorini, G. Grom, L. Shepard, C. G. Stangor, B. Weiner, and T. A. McKay, Patterns of gendered performance differences in large introductory courses at five research universities, *AERA Open* **3** (2017).
- [7] A. Malespina and C. Singh, Gender differences in grades versus grade penalties: Are grade anomalies more detrimental for female physics majors?, *Phys. Rev. Phys. Educ. Res.* **18**, 020127 (2022).
- [8] K. M. Whitcomb, S. Cwik, and C. Singh, Not all disadvantages are equal: Racial/Ethnic minority students have largest disadvantage among demographic groups in both STEM and non-STEM GPA, *AERA Open* **7** (2021).
- [9] X. Chen, STEM attrition: college students' paths into and out of STEM fields. Statistical analysis report. NCES 2014-001, Technical Report, National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C., 2013.
- [10] M. E. Thompson, Grade expectations: The role of first-year grades in predicting the pursuit of STEM majors for first- and continuing-generation students, *J. Higher Educ.* **92**, 961 (2021).
- [11] T. Sanabria and A. Penner, Weeded out? Gendered responses to failing calculus, *Soc. Sci.* **6**, 47 (2017).
- [12] J. R. Cimpian, T. H. Kim, and Z. T. McDermott, Understanding persistent gender gaps in STEM, *Science* **368**, 1317 (2020).
- [13] E. M. Marshman, Z. Y. Kalender, T. Nokes-Malach, C. Schunn, and C. Singh, Female students with A's have similar physics self-efficacy as male students with C's in introductory courses: A cause for alarm?, *Phys. Rev. Phys. Educ. Res.* **14**, 020123 (2018).
- [14] N. Hatfield, N. Brown, C. M. Topaz, and M. Gelfand, Do introductory courses disproportionately drive minoritized students out of STEM pathways?, *PNAS Nexus* **1**, pgac167 (2022).
- [15] American Physical Society, Statistics on Diversity in Physics.
- [16] D. J. Webb, C. A. Paul, and M. K. Chessey, Relative impacts of different grade scales on student success in introductory physics, *Phys. Rev. Phys. Educ. Res.* **16**, 020114 (2020).
- [17] C. A. Paul and D. J. Webb, Percent grade scale amplifies racial or ethnic inequities in introductory physics, *Phys. Rev. Phys. Educ. Res.* **18**, 020103 (2022).
- [18] A. B. Simmons and A. F. Heckler, Grades, grade component weighting, and demographic disparities in introductory physics, *Phys. Rev. Phys. Educ. Res.* **16**, 020125 (2020).
- [19] D. J. Webb and C. A. Paul, Attributing equity gaps to course structure in introductory physics, *Phys. Rev. Phys. Educ. Res.* **19**, 020126 (2023).
- [20] D. J. Webb and C. A. Paul, High-stakes exams inflate a gender gap and contribute to systematic grading errors in one introductory physics series, *Phys. Rev. Phys. Educ. Res.* **21**, 020134 (2025).
- [21] S. Suresh and A. F. Heckler, Evolution of grades and social comparison concern within an introductory physics course, *Phys. Rev. Phys. Educ. Res.* **19**, 010129 (2023).
- [22] S. Cotner, C. J. Ballen, and A. R. Dalby, Can mixed assessment methods make biology classes more equitable?, *PLOS One* **12**, e0189610 (2017).
- [23] M. Levinson, T. Geron, and H. Brighthouse, Conceptions of educational equity, *AERA Open* **8** (2022).
- [24] T. Russo-Tait, Science faculty conceptions of equity and their association to teaching practices, *Sci. Educ.* **107**, 427 (2023).
- [25] I. Rodriguez, E. Brewe, V. Sawtelle, and L. H. Kramer, Impact of equity models and statistical measures on interpretations of educational reform, *Phys. Rev. ST Phys. Educ. Res.* **8**, 020103 (2012).
- [26] L. E. Kost, S. J. Pollock, and N. D. Finkelstein, Characterizing the gender gap in introductory physics, *Phys. Rev. ST Phys. Educ. Res.* **5**, 010101 (2009).
- [27] A. Madsen, S. B. McKagan, and E. C. Sayre, Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap?, *Phys. Rev. ST Phys. Educ. Res.* **9**, 020121 (2013).
- [28] M. Huberth, P. Chen, J. Tritz, and T. A. McKay, Computer-tailored student support in introductory physics, *PLOS One* **10**, e0137001 (2015).
- [29] R. Matz, K. Schulz, E. Hanley, H. Derry, B. Hayward, B. Koester, C. Hayward, and T. McKay, Analyzing the efficacy of ECoach in supporting gateway course success through tailored support, in *LAK21 11th International Learning Analytics and Knowledge Conference, LAK21* (Association for Computing Machinery, New York, NY, USA, 2021), pp. 216–225.
- [30] S. Lonn and B. Koester, Rearchitecting data for researchers: A collaborative model for enabling institutional learning analytics in higher education, *J. Learn. Anal.* **6**, 107 (2019).
- [31] G. A. Morris, P. Walter, S. Skees, and S. Schwartz, Transition matrices: A tool to assess student learning and improve instruction, *Phys. Teach.* **55**, 166 (2017).
- [32] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/ym5f-hs2g> for an analysis of students with missing exam scores, average scores on exams, variations in exam scores, variations in exam letter grades, the Pearson correlations for the numeric scores between

- exams, and results presented by non-binary racial and ethnic categories. It includes Refs. [33, 34].
- [33] M. G. Kendall, A new measure of rank correlation, *Biometrika* **30**, 81 (1938).
 - [34] M. G. Kendall, The treatment of ties in ranking problems, *Biometrika* **33**, 239 (1945).
 - [35] J. Cohen, Statistical power analysis for the behavioral sciences, 2nd ed. (Routledge, New York, 2013).
 - [36] N. T. Young, M. Mills, R. L. Matz, E. F. Bell, and C. Hayward, Exploring how complex multiple-choice questions could contribute to inequity in introductory physics, *PLOS One* **20**, e0323813 (2025).
 - [37] L. Westbrook and A. Saperstein, New categories are not enough: Rethinking the measurement of sex and gender in social surveys, *Gender Soc.* **29**, 534 (2015).
 - [38] C. D'ignazio and L. F. Klein, *Data Feminism* (MIT Press, Cambridge, Massachusetts, 2020).
 - [39] D. Shafer, M. S. Mahmood, and T. Stelzer, Impact of broad categorization on statistical results: How underrepresented minority designation can mask the struggles of both Asian American and African American students, *Phys. Rev. Phys. Educ. Res.* **17**, 010113 (2021).
 - [40] A.-L. Boulesteix and S. Hoffmann, To adjust or not to adjust: It is not the tests performed that count, but how they are reported and interpreted, *BMJ Med.* **3**, e000783 (2024).
 - [41] P. Ramanathan and E. Fernandez, Can early-assignment grades predict final grades in IT courses?, in *2017 ASEE Annual Conference & Exposition Conference Proceedings* (American Society for Engineering Education, Columbus, Ohio, 2017).
 - [42] K. Guidry, Can first exam grades and mathematical preparedness predict success in managerial finance?, *J. Instructional Pedagogies* **23** (2020).
 - [43] H. Li and X. Xia, Grades as signals of comparative advantage: How letter grades affect major choices, *J. Econ. Behav. Organization* **227**, 106717 (2024).
 - [44] L. Hadsell and R. MacDermott, Grade dropping, strategic behavior, and student satisficing, *Am. J. Bus. Educ.* **3**, 57 (2010).
 - [45] E. Sewell, Grade dropping: An empirical analysis, *J. Econ. Educ.* **35**, 24 (2004).
 - [46] R. J. MacDermott, The impact of assessment policy on learning: Replacement exams or grade dropping, *J. Econ. Educ.* **44**, 364 (2013).
 - [47] E. Dabbour, Motivating engineering students by providing two midterm exams and dropping the lower mark, *J. Civ. Eng. Educ.* **147**, 04020009 (2021).
 - [48] G. S. Oettinger, The effect of nonlinear incentives on performance: Evidence from "Econ 101," *Rev. Econ. Stat.* **84**, 509 (2002).
 - [49] J. B. Main and B. Ost, The impact of letter grades on student effort, course selection, and major choice: A regression-discontinuity analysis, *J. Econ. Educ.* **45**, 1 (2014).
 - [50] D. Grant and W. B. Green, Grades as incentives, *Empirical Econ.* **44**, 1563 (2013).
 - [51] T. Gray and J. Bunte, The effect of grades on student performance: Evidence from a quasi-experiment, *Coll. Teach.* **70**, 15 (2022).
 - [52] G. D. Kuh and S. Hu, Unraveling the complexity of the increase in college grades from the mid-1980s to the mid-1990s, *Educ. Eval. Policy Anal.* **21**, 297 (1999).
 - [53] S. Rojstaczer and C. Healy, Grading in American colleges and universities, *Teach. Coll. Rec.* **4**, 1 (2010).
 - [54] C. Jephcote, E. Medland, and S. Lygo-Baker, Grade inflation versus grade improvement: Are our students getting more intelligent?, *Assess. Eval. Higher Educ.* **46**, 547 (2021).
 - [55] A. Evrard, K. Schulz, and C. Hayward, How did you get that A? selectivity's role in rising undergraduate grades at a large public university, in *LAK21: 11th International Learning Analytics and Knowledge Conference, LAK21* (Association for Computing Machinery, New York, NY, USA, 2021), pp. 565–571.
 - [56] G. Kulick and R. Wright, The impact of grading on the curve: A simulation analysis, *Int. J. Scholarship Teach. Learn.* **2** (2008).
 - [57] M. Thulasidas, Statistical moderation: A case study in grading on a curve, in *2021 IEEE International Conference on Engineering, Technology & Education (TALE)*, Wuhan, Hubei Province, China (IEEE, 2021), pp. 734–739.
 - [58] R. S. Bowen and M. M. Cooper, Grading on a curve as a systemic issue of equity in chemistry education, *J. Chem. Educ.* **99**, 185 (2022).
 - [59] A. Malespina and C. Singh, Gender differences in test anxiety and self-efficacy: Why instructors should emphasize low-stakes formative assessments in physics courses, *Eur. J. Phys.* **43**, 035701 (2022).
 - [60] S. Kanim and X. C. Cid, Demographics of physics education research, *Phys. Rev. Phys. Educ. Res.* **16**, 020106 (2020).
 - [61] A. Malespina, F. Seifollahi, and C. Singh, Bioscience students in physics courses with higher test anxiety have lower grades on high-stakes assessments and women report more test anxiety than men, *Educ. Sci.* **14**, 1092 (2024).